DOCUMENT RESUME

ED 392 840 TM 024 622

AUTHOR De Ayala, R. J.; And Others
TITLE An Investigation of the Standard Errors of Expected A Posteriori Ability Estimates.
PUB DATE Apr 95
NOTE 43p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Bayesian Statistics; *Error of Measurement; *Estimation (Mathematics); *Maximum Likelihood Statistics; Monte Carlo Methods
IDENTIFIERS Ability Estimates; *Expected A Posteriori Tests

ABSTRACT

Expected a posteriori has a number of advantages over maximum likelihood estimation or maximum a posteriori (MAP) estimation methods. These include ability estimates (thetas) for all response patterns, less regression towards the mean than MAP ability estimates, and a lower average squared error. R. D. Bock and R. J. Mislevy (1982) state that the posterior standard deviation (PSD theta) is virtually interchangeable with the standard error (SEE). A typical criterion for terminating an adaptive test is when the theta's SEE is equal to or less than a predetermined value. However, if there are conditions in which the PSD theta is not interchangeable with the SEE, then the adaptive test may not be validly terminated. The use of the PSD theta in situations where an examinee is classified on the basis of his or her ability may lead to incorrect classifications if the PSD theta does not agree with the SEE. Results of this Monte Carlo study showed that the use of 10 quadrature points tends to result in PSD thetas which underestimate the observed standard error. The use of 80 quadrature points, given the test's length, is recommended where accurate PSD thetas are required. (Contains 8 figures, 14 tables, and 14 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 392 840

An Investigation of the Standard Errors of Expected A Posteriori Ability Estimates

R.J. De Ayala,

William D. Schafer, and

Monica Sava-Bolesta

University of Maryland

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

RALPH DE AYALA

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Please send correspondence to :

R.J. De Ayala

Measurement, Statistics, and Evaluation

Benjamin Building

University of Maryland

College Park, MD 20742

USA

TM024622

BEST COPY AVAILABLE

An Investigation of the Standard Errors of Expected A Posteriori Ability Estimates

An Investigation of the Standard Errors of Expected A Posteriori Ability Estimates

ABSTRACT

Expected a posteriori (EAP) has a number of advantages over maximum likelihood estimation (MLE) or maximum a posteriori (MAP) estimation methods. These include ability estimates ($\hat{\theta}$s) for all response patterns including zero and perfect score patterns, less regression towards the mean than MAP ability estimates, and an average squared error that is less than that for MAP and MLE $\hat{\theta}$s. Bock and Mislevy (1982) state that the posterior standard deviation (PSD($\hat{\theta}$)) is virtually interchangeable with the standard error (SEE). A typical criterion for terminating an adaptive test is when the $\hat{\theta}$'s SEE is equal to or less than a predetermined value. However, if there are conditions in which the PSD($\hat{\theta}$) is not interchangeable with the SEE, then the adaptive test may not be validly terminated. Moreover, in applications where an examinee must be classified on the basis of his/her ability estimate (e.g., as a master versus nonmaster) one typically creates a confidence interval about the examinee's ability estimate using the $\hat{\theta}$'s SEE. The use of the PSD($\hat{\theta}$) in these situations may lead to incorrect classifications if the PSD($\hat{\theta}$) does not agree with the SEE. Results of this Monte Carlo study showed that the use of 10 quadrature points tends to result in PSD($\hat{\theta}$)s which underestimate the observed standard error. The use of 80 quadrature points, given the test's length (possibly 2 $*\sqrt{\text{test length}}$ quadrature points under certain conditions), is recommended where accurate PSD($\hat{\theta}$)s are required.

Item response theory (IRT) has emerged as a popular approach for solving various measurement problems, such as test design, test equating, and computerized adaptive testing (CAT), and IRT techniques are becoming more common in practical testing situations. For example, certification boards such as the American Society of Clinical Pathologists have established an IRT-based CAT system for certification (Bergstrom & Lunz, 1991). Unlike the conventional paper-and-pencil test in which an examinee, regardless of ability, is administered all test items, CAT is a procedure for administering tests that are individually tailored for each examinee. Advantages of IRT-based CAT over paper-and-pencil testing have been well documented (e.g., Wainer, 1990; Weiss, 1982). Although not necessary, a CAT system typically uses an IRT model in combination with item characteristics to estimate the examinee's ability.

Ability estimation in CAT has typically used one of three methods: maximum likelihood estimation (MLE) or Bayesian approaches such as maximum a posteriori (MAP, also known as Bayes Modal Estimate) and expected a posteriori (EAP, also known as Bayes Mean Estimate). The former two algorithms are iterative techniques, while EAP is noniterative and is based on numerical quadrature methods. Because it is noniterative (and efficient) it is potentially faster than either MLE or MAP in ability estimation. The obvious implication of EAP's efficient estimation for CAT is the transparency (as far as the examinee is concerned) of estimating the examinee's ability in real time, particularly with more complicated IRT models (e.g., polytomous IRT models). Moreover, unlike MLE ability estimates, EAP ability estimates may be obtained for all response patterns, including zero and perfect score patterns (Mislevy & Stocking, 1989). While MAP ability estimates also exist for all response patterns, they suffer from greater regression towards the mean than do the EAP estimates (Bock & Mislevy, 1982; Mislevy & Bock, 1982). Moreover, in the early stages of an adaptive test the EAP estimate is more stable than the MAP estimate and the average squared error for EAP estimates over the population of ability is less than that for MAP and MLE ability estimates (Bock & Mislevy, 1982). From an implementation perspective, an additional advantage is the simplicity of the mathematics required for deriving the computational forms for ability estimation with polytomous IRT models.

The EAP estimate (Bock & Mislevy, 1982) of an examinee's ability, θ, after n items have been administered is given by

$$\hat{\theta}_n = \frac{\sum_{k=1}^{q} X_k \, L_n(X_k) \, A(X_k)}{\sum_{k=1}^{q} L_n(X_k) \, A(X_k)} \qquad (1)$$

and its posterior standard deviation is

$$PSD(\hat{\theta}) = \sqrt{\frac{\sum_{k=1}^{q} (X_k - \hat{\theta}_n)^2 L_n(X_k) A(X_k)}{\sum_{k=1}^{q} L_n(X_k) A(X_k)}} \qquad (2)$$

where X_k is one of q quadrature points, $A(X_k)$ is the quadrature weight associated with X_k, and $L_n(X_k)$ is the likelihood function of X_k given the response pattern $\{x_1, x_2, ..., x_n\}$. For example, if the probability of a correct response by an individual with ability θ to a dichotomously score item i with location b_i is given by the one-parameter logistic (1PL) model

$$p(x_i = 1|\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \qquad (3)$$

then the likelihood of θ given the response pattern $\{x_1, x_2, ..., x_n\}$ is

$$L_n(\theta) = \prod_{i=1}^{n} p(x_i = 1|\theta)^{x_i}(1 - p(x_i = 1|\theta))^{(1 - x_i)} \qquad (4)$$

The X_ks and $A(X_k)$s may be obtained from tables provided by Stroud and Secrest (1966) for approximating the Gaussian error function. The Stroud and Secrest Gauss-Hermite X_ks and $A(X_k)$s must be multiplied by $\sqrt{2}$ and $\frac{1}{\sqrt{\pi}}$ (Bock & Lieberman, 1970), respectively, in order to place them on the normal function scale. However, programs, such as BILOG (Mislevy & Bock, 1982), do not use the Stroud and Secrest values for EAP ability estimation; neither BILOG nor MULTILOG (Thissen, 1988) use these values for obtaining item parameter estimates via marginal maximum likelihood estimation (MMLE). Rather, a specified range of the θ continuum (e.g., -4.0 to 4.0) is divided into q equidistant discrete points (these points serve as the X_ks) and the standard unit normal probability density is computed at each of the q points. The probability density at X_k multiplied by the difference between successive quadrature points (e.g., $X_k - X_{k+1}$) is the quadrature weight $A(X_k)$. Because of the symmetric nature of the discrete prior distribution the $A(X_k)$s only need to be calculated for the X_ks ≤ 0. (Seong (1990a) refers to this method as the "Mislevy histogram" technique, although it is probably more accurate to refer to it as the Mislevy "vertical line graph" method to emphasize the discrete nature of the prior distribution.) Seong (1990a) has compared the item and ability parameter estimates obtained by using this latter technique with those obtained by the Stroud and Secrest values. Seong found that when a large number of quadrature points were used (e.g., 30 or 40) the two methods estimated item and ability parameters equally well, but when a small number of quadrature points were specified (e.g., 10), the Mislevy histogram solution estimated item and ability parameters more accurately than the Gauss-Hermite quadrature formula. It should be noted that Bock and Mislevy (1982) state that the Gauss-Hermite values do not include

the likelihood functions found in adaptive testing. Moreover, the X_ks and $A(X_k)$s must satisfy the constraints that $\sum A(X_k) = 1.0$, $\sum X_k A(X_k) = 0.0$, and $\sum X_k^2 A(X_k) = 1.0$.

Bock and Mislevy's (1982) work showed that EAP produces reasonably accurate ability estimates. Originally, Bock and Mislevy presented EAP for use in adaptive testing, however in the calibration program BILOG (Mislevy & Bock, 1982) it is the default ability estimation approach. In adaptive testing the $PSD(\hat{\theta})$ plays the same role as the MLE's standard error (Bock & Mislevy, 1982). That is, after 20 items the likelihood function and the posterior distribution are nearly identical and the $PSD(\hat{\theta})$ is virtually interchangeable with the standard error (Bock & Mislevy, 1982); this interchangeability is reflected in the fact that the $PSD(\hat{\theta})$s are labeled as standard errors in the BILOG EAP output. For consistency with and on the basis of Bock and Mislevy (1982), the EAP $PSD(\hat{\theta})$ will be referred to as if it were a standard error and will be labeled as EAP SEE in the following.

A number of studies have investigated the effects of various factors on MMLE item parameter estimation (e.g., Drasgow, 1989; Harwell & Janosky, 1991; Zwinderman & van der Wollenberg, 1990). Seong (1990b) evaluated both item parameter estimation and EAP ability estimation. With respect to EAP $\hat{\theta}$s, Seong found that increasing the number of quadrature points from 10 to 20 produced more accurate $\hat{\theta}$s, regardless of sample size and appropriateness of the prior distribution (i.e., normal, positively and negatively skewed). Because abilities are estimated independently of one another it is not surprising that sample size did not have a significant effect on the accuracy of EAP $\hat{\theta}$s. Because of the breadth of Seong's study, the EAP estimation findings were limited. For instance, test length should affect ability estimation, but was held fixed at 45 items in Seong's study. Moreover, Seong studied the accuracy of the $\hat{\theta}$s in terms of root mean square error (not EAP SEE), but in applications where an examinee must be classified on the basis of his/her ability estimate (e.g., as a master versus nonmaster) one typically creates a confidence interval about the examinee's ability estimate using the $\hat{\theta}$'s SEE. As an example, in the American Society of Clinical Pathologists' CAT pathologists are presented an adaptive certification test. If the confidence interval for an examinee falls either completely above or completely below the cut point, then the examinee may be classified as a master (i.e., certified) or a nonmaster, respectively. If the confidence interval spans the cut point, then additional information is needed (e.g., more test questions could be asked). The use of confidence intervals incorporates our uncertainty about the ability estimate. It should also be noted that in addition to using the EAP $PSD(\hat{\theta})$ as if it was a standard error, the $PSD(\hat{\theta})$ calculated by (2) is actually an estimate or an approximation and its use for forming confidence intervals may be problematic if the EAP SEE is not accurate. Moreover, a typical criterion for terminating an adaptive test is when the $\hat{\theta}$'s SEE is equal to or less than a predetermined value. If the EAP SEE is not accurate, then the adaptive test

may not be validly terminated. For these reasons this study was primarily concerned with the validity of the EAP SEEs.

Because EAP is based on numerical quadrature methods it requires the specification of a number of factors, such as type of prior distribution and the number of quadrature points to use in estimation. Each of these factors as well as the test length and the form of the examinees' latent distribution may affect the accuracy of the EAP ability estimate and the EAP SEE. This study investigated the effects of the number of quadrature points (10, $2*\sqrt{\text{test length}}$, and 80), test length (61 and 122 items), latent ability distribution (bimodal, normal, positively skewed, and uniform), and the form of the prior ability distribution (normal and uniform) on the EAP SEEs. The $2*\sqrt{\text{test length}}$ and 80 number of quadrature point levels were chosen because $2*\sqrt{\text{test length}}$ is the default value in BILOG for EAP estimation (a normal prior is also default) and according to Bock and Mislevy (1982, p. 433) "In applications to real populations, perhaps 80 quadrature points between ± 4.0 standard deviations should be available to insure precision down to PSD = 0.2" (although for their simulation they used 21 quadrature points). A bimodal latent ability distribution was used to simulate an examinee population that consists of masters and nonmasters, and the rationale for the test lengths is presented below.

METHOD

Program: A program was written for generating simulees, generating the responses for each simulee, performing ability estimation for each simulee, and compiling various summary statistics for each simulee as well as across simulees.

Data. For each of the 4 latent distributions, 100 simulees were sampled from the appropriate θ distribution. Then for each simulee at each combination of test length, prior distribution and quadrature points, the process of administering a simulated test, as described below, was repeated 1000 times. The standard unit normal curve was used as the θ distribution for the normal condition, a beta distribution ($v_1 = 1.25$, $v_2 = 10$) was used to produce the positively skewed θ distribution (skew = 1.14), and the uniform θ distribution was restricted to the range $-3.0 \leq \theta \leq 3.0$. The bimodal θ distribution was obtained by generating one-half the sample's simulees from a beta distribution with $v_1 = 1.25$ and $v_2 = 10$ and one-half from a beta distribution with v_1 and v_2 transposed. Each latent ability distribution had a unique seed for generating its simulees.

A sixty-one item pool was generated to have uniform difficulty parameters (b) in the range $-3.0 \leq b \leq 3.0$ in 0.1 logit increments (i.e., $b_1 = -3.0$, $b_2 = -2.9$, etc.). The discrimination (a) and the pseudo-guessing (c) parameters were set at 1.0 and 0.0, respectively. The use of these values for a and c is discussed below. The 122-item test consisted of the 61-item test replicated and therefore the 122-item test information function was twice that of the 61-item test.

For each simulee, responses were generated using the appropriate item parameters, and the simulee's θ to calculate the probability of obtaining the item correct according to the 1PL model. This probability was compared to a random number obtained from a uniform distribution [0,1]. If the probability was greater than the random number then the simulee's response was 1 (i.e., correct), otherwise the simulee's response was incorrect and coded as 0.

After the simulee had been administered a test of the appropriate length an EAP $\hat{\theta}$ and its EAP SEE were obtained using the appropriate prior distribution and number of quadrature points. This process was repeated 1000 times for each simulee (i.e., there were 1000 $\hat{\theta}$s for each of the 100 θ in each of the 48 cells in the design).

Estimation. EAP ability estimates were calculated according to (1) and the EAP SEE was obtained according to (2). For the three levels of the number of quadrature factor (10, $2*\sqrt{\text{test length}}$, and 80) the X_ks and $A(X_k)$s were determined using the Mislevy "vertical line graph" method described above for the range $-4.0 \leq \theta \leq 4.0$. For the 61-item test $2*\sqrt{\text{test length}} = 16$ and for the 122-item test $2*\sqrt{\text{test length}} = 23$.

Analyses: In addition to obtaining the EAP SEE of $\hat{\theta}$, the standard deviation of the 1000 $\hat{\theta}$s (i.e., the empirical SEE) for a given θ was calculated. The basic design of the study was a four-way repeated measures design with the difference between the empirical and EAP SEEs (i.e., $SEE_{empirical} - SEE_{EAP}$) as the dependent variable, latent ability distribution as the between subjects factor, and test length, type of prior distribution, and number of quadrature points as the within subjects factors.

In addition to calculating the empirical SEE, 68% and 95% confidence intervals (CIs) based on the EAP SEE were calculated and the number of times the 68% and 95% CIs contained θ were counted (CI68% and CI95%, respectively). Analysis of the CIs involved calculating the difference between the number of times a given CI contained θ and the number of times the CI was expected to contain θ (i.e., diff68% = CI68% - 680 and diff95% = CI95% - 950). The analyses of diff68% and diff95% were treated separately. Diff68% and diff95% were used as the dependent variable in a four-way repeated measures analysis with test length, type of prior distribution, and number of quadrature points as the within subjects factors and latent ability distribution as the between subjects factor.

The accuracy of ability estimation was assessed by root mean square error (RMSE) and Bias. RMSE and Bias were calculated according to:

$$RMSE(\theta) = \sqrt{\frac{\sum (\hat{\theta}_k - \theta)^2}{N_k}} \qquad (5)$$

$$Bias(\theta) = \frac{\sum (\hat{\theta}_k - \theta)}{N_k} , \qquad (6)$$

where $\hat{\theta}_k$ is the ability estimate for simulee k with latent ability θ, and N is the number of replications for simulee k (i.e., N_k = 1000). RMSE was used as the dependent variable in four-way repeated measures analysis with within subjects factors of test length and number of quadrature points, type of prior distribution, and latent ability distribution as the between subjects factor. The analysis of Bias was treated similarly. Descriptive statistics were calculated on the θs and $\hat{\theta}$s as well as on the EAP and empirical SEEs, the difference between SEEs, CI68% and CI95%. Fidelity coefficients ($r_{\theta\hat{\theta}}$) were obtained.

To summarize, the effect of the four factors (latent ability distribution, prior distribution, test length, and number of quadrature points) on the EAP and empirical SEEs was investigated using a four-way repeated measures design for SEE. The two CIs were each analyzed using a four-way repeated measures analysis with diff68% and diff95% as the dependent variables. Accuracy of ability estimation was assessed using a four-way repeated measures analysis with RMSE and Bias as the dependent variables. Because of its relaxed assumptions a multivariate approach was used for all repeated measures analyses.

RESULTS

Tables 1 and 2 contain the descriptive statistics on the θs and $\hat{\theta}$s as well as the $r_{\theta\hat{\theta}}$. As can be seen from Table 1, increasing the number of quadrature points from 10 to $2*\sqrt{\text{test length}}$ and to 80, resulted in the mean and median $\hat{\theta}$ becoming more similar to the mean and median θ, respectively, regardless of latent distribution, prior distribution, and test length level. Table 2 shows that the $r_{\theta\hat{\theta}}$ also increased as the number of quadrature points increased from 10 to 80 nodes regardless of latent distribution, prior distribution, and test length level. However, these increases in $r_{\theta\hat{\theta}}$ may not be considered meaningful by some because of the magnitude of the $r_{\theta\hat{\theta}}$ at the 10 quadrature point level.

Insert Tables 1 and 2 about here

Descriptive statistics on the empirical and EAP SEEs are presented in Table 3. This table shows that increasing the number of quadrature points led to a decrease in the mean empirical SEEs regardless of test length, prior distribution, and latent distribution. As would be expected, doubling the test length led to a decrease in the average SEEs for all levels of the number of quadrature points factor. Furthermore, for a given latent and prior distribution the mean empirical SEE for the 10 quadrature point level/122-item test length was, typically, approximately the same size as the average SEE for the 16 quadrature point level/61-item test length and in certain conditions less than those for the 80 quadrature point level at the shorter test length. In general, as the number of quadrature points increased the average empirical SEEs decreased. In contrast, the EAP SEEs showed the opposite pattern with respect to increasing the number of quadrature points. Specifically, increasing the number of quadrature points led to an

increase in the mean EAP SEEs. A comparison of the EAP and empirical SEEs shows that, regardless of test length, latent and prior distribution, the mean EAP SEEs for the $2*\sqrt{\text{test length}}$ and 80 quadrature point levels had a tendency to be in good agreement with the mean empirical SEEs. However, the average EAP SEEs tended to underestimate the mean empirical SEEs when 10 quadrature points was used for estimation, but as the number of quadrature points increased the EAP SEEs and empirical SEEs came into closer agreement. As was the case with the empirical SEEs, doubling the test length had the expected effect of decreasing the average EAP SEEs. The discrepancy between the EAP and empirical SEEs was greatest for the positively skewed latent ability distribution.

Insert Table 3 about here

The descriptive statistics on CI68% and CI95% are presented in Table 4. Given the SEE results it is not surprising that the CI68% and CI95% were affected by the number of quadrature points. It is only when 80 quadrature points were used for ability estimation that the CI68% and CI95% approximated their expected values of 680 and 950, respectively, regardless of test length, prior and latent distributions.

Insert Table 4 about here

Table 5 contains the descriptive statistics on RMSE(θ) and Bias(θ). For the normal, positively skewed, and uniform latent distributions increasing the number of quadrature points from 10 to 80 nodes led to more accurate $\hat{\theta}$ on average. However, for the bimodal condition there was a slight increase in the mean RMSE(θ) as the number of nodes increased from $2*\sqrt{\text{test length}}$ to 80. For a given number of quadrature points and independent of the latent and prior distributions, doubling the test length resulted in a decrease in the average RMSE(θ).

The mean Bias(θ) values tended to about 0.0 (range of -0.074 to 0.021) and inspection of the corresponding histograms showed that these distributions tended to be somewhat unimodal and symmetric about 0.0. There were five instances of bimodal distributions (3 associated with the bimodal and 2 with the uniform latent ability distributions) and these occurred with the use of a normal prior. Table 5 also shows that there was a slight underestimation Bias(θ) for the bimodal and positive skew θ distributions and in general, a slight overestimation for the normal and uniform latent ability distributions. The standard deviations of Bias(θ) were correspondingly small and decreased with increasing number of quadrature points indicating that the average Bias(θ) value was a "typical" Bias(θ) value and not atypically low because of the compensation that takes place in its calculation (see (6)).

Insert Table 5 about here

The repeated measures analyses on the SEE difference are presented in Table 6. As can be seen the magnitude of the difference between the EAP and empirical SEEs was affected by type of the latent distribution, the test length, and the type of prior distribution as well as the number of quadrature points used in estimation. Post hoc analyses on the SEE difference (Table 7) showed that for the bimodal and uniform latent distribution conditions increasing the number of quadrature points from $2 * \sqrt{\text{test length}}$ to 80 did not result in a significant improvement in the agreement between the EAP and empirical SEEs. This was also true for the normal and positively skewed θ distributions, but only for the 122-item test. However, the use of the 61-item test with these two θ distributions showed that increasing the number of quadrature points from $2 * \sqrt{\text{test length}}$ to 80 did result in a significant improvement in the agreement between the EAP and empirical SEEs. It should be noted that the disagreement between EAP and empirical SEEs for the normal and positively skewed θ distributions using a 61-item test with $2 * \sqrt{\text{test length}}$ quadrature points is less than 0.044 (Table 3).

Insert Tables 6 and 7 about here

Table 7 also shows that for EAP estimation based on 80 quadrature points and a uniform prior distribution doubling the test length did not result in a significant improvement in the agreement between EAP and empirical SEEs. This pattern held for the normal prior except for the uniform latent distribution condition where the test statistic was marginally significant. The use of $2 * \sqrt{\text{test length}}$ quadrature points, a uniform prior, and 122-item test produced significantly greater agreement between the EAP and empirical SEEs for all latent ability distributions. There was not as clear a pattern for the other conditions and while it was expected that when the prior distribution matched the latent ability distribution there would be better agreement between the EAP and empirical SEEs than when there was a mismatch, this pattern did not appear. It should be noted that the magnitude of the SEE differences were comparatively small for the $2 * \sqrt{\text{test length}}$ and 80 quadrature point conditions (i.e., discrepancies in the hundreds and thousandths) and only at 10 quadrature points were these discrepancies occurring at the first decimal place. In this regard as well as with respect to the power of the tests, some of the significant post hocs may not be considered meaningful by some.

Figure 1 contains the test length by quadrature points by prior distribution interaction plot for each latent ability distribution. The plots clearly show (a) the convergence of empirical and EAP SEEs as the number of quadrature points increased; (b) for all θ distributions the SEEs for the 122-item test were less than those for a test half as long for a given quadrature point, prior distribution and type of SEE (i.e., empirical or EAP) level; and (c) for a given quadrature point level and for a SEE type, the use of a uniform prior resulted in larger values than the use of a

normal prior at the 61-item test length (this difference appeared to disappear at the 122-item test length).

Insert Figure 1 about here

Table 8 contains the repeated measures analyses for the confidence intervals. For diff68% all first-order interactions and the latent ability distribution by test length by number of quadrature points interaction were significant, whereas for diff95% the four-way interaction of latent ability distribution, test length, number of quadrature points, and type of prior distribution was significant.

Insert Table 8 about here

Post hoc analysis of the effect of type of prior distribution used in estimation on diff68% (Table 9) showed that for a given test length the use of a uniform prior distribution, rather than a normal prior, led to significantly better agreement between the average number of 68% CIs containing θ and their expected value of 680. However, for a given prior distribution doubling the test length led to a significant increase in the mean number of 68% CIs not containing θ. Inspection of Table 4 showed that these significant results were associated with poorer performance (i.e., lack of agreement between the number of 68% CIs containing θ approaching their expected value of 680) at the 10 quadrature point level for the 122-item test than at the 61-item test length, regardless of the prior and latent ability distribution. When the number of quadrature points is increased from 10 to $2 * \sqrt{\text{test length}}$ or greater, then doubling the test length produces better agreement between the number of 68% CIs containing θ approaching their expected value of 680 at all levels of prior and latent ability distribution. Moreover, although for a given prior distribution increasing the number of quadrature points led to significant improvement, only when 10 quadrature points were used for estimation was the choice of prior distribution relevant. For instance, the use of 10 quadrature points resulted in significantly better agreement between the average number of 68% CIs containing θ and their expected value of 680 when a uniform prior distribution was used instead of a normal distribution. However, when 80 quadrature points were used for EAP estimation the mean diff68% when a normal prior was used was -0.808 and for a uniform prior it was 0.518 and the choice of prior was irrelevant. While at the $2 * \sqrt{\text{test length}}$ level there was no significant difference for type of prior distribution, there were, on average, 59.71 fewer CI68% not containing θ than would be expected when a normal prior was used and when a uniform prior was used the mean diff68% was -55.80. Therefore, only at the 80 quadrature point level was the number of CI68% containing θ approaching the expected value of 680.

Insert Tables 9 and 10 about here

Analysis of the CI95% (Table 10) showed that increasing the number of quadrature points from 10 to 2 *$\sqrt{\text{test length}}$ resulted in significantly more 95% CIs approaching their expected value of 950 regardless of test length, type of prior distribution, and θ distribution. In addition, increasing the number of quadrature points from 2 *$\sqrt{\text{test length}}$ to 80 led to a significant reduction in the mean diff95% for the 61-item test with the use of a normal prior for all latent ability distributions. This was also true when a uniform prior was used with a 61-item test and when the θ distributions were normal or positively skewed. While there was not a significant difference between the increase from 2 *$\sqrt{\text{test length}}$ to 80 quadrature points for certain conditions (e.g., uniform θ distribution, 122-item test length), a comparison with Table 4 showed that the magnitude of the mean difference for these nonsignificant cells was, at most, 1.9 (the uniform θ distribution, uniform prior, 122-item test length cell). That is, for these nonsignificant cells and when 80 quadrature points were used for estimation there were, on average, 1.9 95% CIs that did not contain θ and overall there were at most 3.6 95% CIs that did not cover the parameter. Therefore, while the difference between 2 *$\sqrt{\text{test length}}$ and 80 quadrature points may not be significant, in practice the number of CIs which contain θ when 80 as oppose to 2 *$\sqrt{\text{test length}}$ quadrature points were used for estimation may be considered meaningful by some. For CI68% and using 80 quadrature points for estimation there were at most, on average, 7.2 68% CIs that did not include θ (Table 4).

Table 11 contains the repeated measures analyses for RMSE(θ) and Bias(θ). These results showed that the second-order interactions for RMSE(θ) were significant, while Bias(θ) was affected by the interaction of θ distribution, test length, number of quadrature points, and type of prior distribution used. Post hoc analyses for RMSE(θ) (Table 12) revealed that there was not a significant interaction between type of prior distribution and the number of quadrature points within latent ability distribution. In general, increasing the number of quadrature points from 10 to 2 *$\sqrt{\text{test length}}$ and from 10 to 80 led to a significant reduction in the mean RMSE(θ), but increasing from 2 *$\sqrt{\text{test length}}$ to 80 quadrature points did not result in significantly more accurate $\hat{\theta}$s, regardless of θ distribution (cf., Table 5). The use of a uniform prior instead of a normal prior led to a significant increase in the average RMSE(θ), however, the magnitude of these increases may not be considered meaningful by some individuals (e.g., for the bimodal, normal, positive skew, and uniform θ distributions the mean RMSE(θ) were 0.2265 (normal) vs 0.2334 (uniform), 0.2554 (normal) vs 0.2622 (uniform), 0.2731 (normal) vs 0.2806 (uniform), and 0.2485 (normal) vs 0.2544 (uniform), respectively).

Insert Table 11 about here

The analysis of the quadrature points by test length within θ distribution interaction showed a significant quadrature points by test length interaction. For all levels of the quadrature points factor increasing the test length from 61 to 122 items produced significantly more accurate $\hat{\theta}$s, regardless of latent ability distribution. For all latent ability distributions, except for the positive skew θ distribution, increasing the number of quadrature points from 10 to 2 *$\sqrt{\text{test length}}$ and from 10 to 80 led to a significant reduction in the mean RMSE(θ), but increasing from 2 *$\sqrt{\text{test length}}$ to 80 quadrature points did not result in significantly more accurate $\hat{\theta}$s, regardless of θ distribution and test length.

Insert Table 12 about here

Within latent ability distribution there was not a significant test length by type of prior distribution interaction. As was the case with the quadrature points by test length within θ distribution interaction, doubling the test length led to a significant reduction in the average RMSE(θ) for all θ distributions. Furthermore and similar to the prior distribution by the number of quadrature points within latent ability distribution interaction, the use of a uniform prior instead of a normal prior led to a significant increase in the average RMSE(θ), regardless of latent ability distribution.

There was a significant number of quadrature points by test length interaction for both normal and uniform prior distributions. Increasing the test length from 61 to 122 items produced a significant reduction in RMSE(θ) for all levels of the number of quadrature points factor, regardless of type of prior distribution used in ability estimation. For both types of prior distributions, increasing the number of quadrature points from 10 to 2 *$\sqrt{\text{test length}}$ and from 10 to 80 led to a significant reduction in the mean RMSE(θ), but increasing from 2 *$\sqrt{\text{test length}}$ to 80 quadrature points did not result in significantly more accurate $\hat{\theta}$s.

Post hoc analyses of Bias(θ) are presented in Table 13. As can be seen all significant differences amongst the levels of the number of quadrature points factor occurred when a 61-item test was used and were reflective of a reduction in the average Bias(θ) at the larger number of quadrature points level from the mean Bias(θ) at the lower number of quadrature points level (cf., Table 5). Similarly, the significant differences between the 61- and 122-item tests were produced by Bias(θ) for the 122-item test being less than that for the 61-item test.

Insert Table 13 about here

CONCLUSIONS AND DISCUSSION

While it may be argued by some that varying *a* and *c* is more realistic with respect to actual testing situations, this study used a 1PL model because it was considered to avoid a number of confounding issues and needlessly complicate the study. A thought-experiment may be sufficient

to consider what may occur using models with varying item discrimination (*a*s) and/or the pseudo-guessing parameter (*c*s). If *a* is allowed to increase from the study's value of 1.0, then given the inverse relationship between information (I(θ)) and SEE the EAP SEEs would become smaller than those obtained in this study. However, because of the greater information available for ability estimation, the $\hat{\theta}$s would become more stable and the empirical SEE would also decrease. In short, the discrepancy between the EAP and empirical SEEs at 10 quadrature points would still exist. The use of items with low *a*s is not considered meaningful because in practice items with low *a*s are not considered desirable (i.e., most psychometrician prefer to use items which discriminate well and to increase I(θ) rather than to decrease it). Using slightly less informative items than used in the study, say $0.8 \leq a < 1.0$, would increase the EAP SEE. However, these same items would make the $\hat{\theta}$s comparatively less accurate and thereby increase the empirical SEEs. The discrepancy between the EAP and empirical SEEs would still remain.

Allowing *c* to increase from the study's value of 0.0 would have a similar impact. When $c > 0.0$ the location of maximum I(θ) simply shifts to be higher than the item's difficulty value (*b*) and lowers the amount of information available for estimation. Therefore, with increasing *c* the variance and standard error of estimation increase. There are two possible scenarios with scenario 1 requiring an assumption. Scenario 1 requires one to assume that by increasing *c* and thereby decreasing the information available for ability estimation it is possible to still obtain reasonably stable and accurate $\hat{\theta}$s (and not increase the empirical SEE). If this is true, then conceivably there is a value of $c > 0.0$ that will sufficiently increase the EAP SEE so that it agrees with the empirical SEE. That is, the goal is to construct a test using items that examinees have a large probability of correctly answering *without knowing the correct answers* (i.e., guessing) and still obtain *accurate* $\hat{\theta}$s for those examinees. Scenario 2 is that increasing *c* and thereby decreasing the information available for ability estimation results in unstable and inaccurate $\hat{\theta}$s. This instability and inaccuracy is reflected in a larger empirical SEE than would be obtained if $c = 0.0$. Therefore, there is no *c* which will sufficiently increase the EAP SEE so that it agrees with the empirical SEE because as *c* increases so does the empirical SEE. It is this latter issue which also addresses the use of "reasonable" *c*s of say, less than 0.25.

To summarize the results of our thought experiment, any nonzero *c* or a value of $a < 1.0$ will increase the empirical and EAP SEEs. Increasing *a* will decrease the EAP and empirical SEEs. In all cases the discrepancy between the EAP and empirical SEEs that was observed at 10 quadrature points will continue to exist.

As mentioned above, Bock and Mislevy (1982) state that the PSD($\hat{\theta}$) is virtually interchangeable with the standard error after about 20 items. Part of the support for this statement comes from their adaptive test simulation results which were based on the use of 21 quadrature points for estimation. This study showed that considering the PSD($\hat{\theta}$) to be

interchangeable with the standard error is questionable even with 122 items if the number of quadrature points is 10; given the trend in the data (see Figure 1) this is probably also true for less than 10 quadrature points. As the number of quadrature points increase it appears that considering $PSD(\hat{\theta})$ to be interchangeable with the standard error is reasonable. For example, given that $RMSE = \sqrt{SEE^2 + Bias(\theta)^2}$, the agreement between the observed mean $RMSE(\theta)$ with the mean RMSEs based on the EAP and empirical SEEs was assessed (Table 14). As can be seen, when the number of quadrature points is 80 there is very good agreement between the observed mean $RMSE(\theta)$ and the RMSEs calculated on the basis of either the EAP SEE or the empirical SEE.

Insert Table 14 about here

This studied showed that when the purpose of assessment is to rank-order examinees in terms of ability, the use of 10 quadrature points provides very good agreement (i.e., $r_{\theta\hat{\theta}}$) between the EAP $\hat{\theta}$s and their corresponding θs for symmetric distributions. If there is reason to suspect that the latent ability distribution is skewed, then the use of $2 * \sqrt{\text{test length}}$ quadrature points may be called for. More accurate $\hat{\theta}$s (i.e., in terms of $RMSE(\theta)$ and $Bias(\theta)$) may be obtained by increasing the test length as well as the number of quadrature points. Furthermore, Table 5 showed that for a fixed test length the accuracy (mean $RMSE(\theta)$) may be increased simply by increasing the number of quadrature points from 10 to 80. For example, the use of 80 quadrature points with a 61-item test produced $RMSE(\theta)$s that were less than those of a test twice as long, but using 10 quadrature points for estimation.

Given the SEE difference, the diff68% and diff95%, and the $RMSE = \sqrt{SEE^2 + Bias(\theta)^2}$ relationship analyses, it appears that the use of 10 quadrature points tends to result in EAP $SEE(\hat{\theta})$s which underestimate the observed standard error. These SEEs give the false impression that the $\hat{\theta}$ is being estimated more accurately than, in fact, it is. Creation of confidence intervals will be erroneously narrower than what they should be and classification decisions based on such CIs will potentially be incorrect. For instance, examinees may be classified as masters (e.g., certified) because their (erroneously narrow) CIs fall above the standard. In these applications it is necessary to increase the number of quadrature points used in EAP estimation. A conservative approach would be to use 80 quadrature points because, overall, this level provided the best agreement between the CIs and their expected values. Clearly, there are situations where the use of $2 * \sqrt{\text{test length}}$ quadrature points may be reasonable given the test's length, the type of prior distribution used, and knowledge of θ's distribution.

When a CAT using EAP ability estimation is terminated using the standard error criterion, it appears necessary to use about 80 quadrature points if the adaptive test will be validly terminated, regardless of latent θ distribution. This is also true if the EAP SEE will be used to

estimate the reliability coefficient; Bock & Mislevy (1982) state that 1 - $PSD(\hat{\theta})^2$ is the reliability coefficient for the EAP $\hat{\theta}$. If it is reasonable to assume a bimodal or uniform θ distribution, then the use 2 $*\sqrt{\text{test length}}$ quadrature points with a normal prior distribution appears to be sufficient for accurate EAP SEEs. However, because of the interaction between test length, number of quadrature points, and EAP SEE, shorter length tests may require greater than 2 $*\sqrt{\text{test length}}$ number of quadrature points to obtain accurate EAP SEEs. Given the observed $r_{\theta\hat{\theta}}$s with 10 quadrature points it may be permissible to use 10 quadrature points in nonadaptive testing situations if the EAP SEE will not be used.

REFERENCES

Bergstrom, B. & Lunz, M. (1991, April). *Confidence in pass/fail decisions for computer-adaptive and paper-and-pencil examinations.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Bock, R.D., & Lieberman, M. (1970). *Fitting a response model for n dichotomously scored items. Psychometrika, 35,* 179-198.

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431-444.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter model. *Applied Psychological Measurement, 13,* 77-90.

Harwell, M.R., & Janosky, J.E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15,* 279-291.

Mislevy, R.J. & Bock, R.D. (1982). *BILOG, maximum likelihood item analysis and test scoring: Logistic model.* Mooresville, IN: Scientific Software, Inc.

Mislevy, R.J., & Stocking, M.L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13,* 57-75.

Seong, Tae-Je. (1990a, April). *Validity of using two numerical analysis techniques to estimate item and ability parameters via MMLE: Gauss-Hermite quadrature formula and Mislevy's histogram solution.* Paper presented at the meeting of the National Council of Measurement in Education, Boston.

Seong, Tae-Je. (1990b). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14,* 299-311.

Stroud, A.H., & Secrest, D. (1966). *Gaussian quadrature formulas.* Englewood Cliffs, NJ: Prentice-Hall.

Thissen, D.J. (1988). *MULTILOG-User's Guide.* Scientific Software, Inc. Mooresville, IN.

Wainer, H. (1990). *Computerized adaptive testing: A primer.* Hillsdale: Lawrence Erlbaum Associates.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473-492.

Zwinderman, A.H., & van der Wollenberg, A.L. (1990). Robustness of marginal maximum likelihood estimation in the Rasch model. *Applied Psychological Measurement, 14,* 73-81.

Table 1: Descriptive statistics on $\hat{\theta}$s and θs.

			Test Length						
			61			122			
Latent θ			Quadrature Points			Quadrature Points			
Distribution	Prior		10	16	80	10	23	80	θ
Bimodal	Normal	Mean	-0.026	-0.011	-0.001	-0.019	-0.001	-0.001	0.000
		Median	-0.619	-0.405	-0.296	-0.729	-0.329	-0.314	-0.308
		SD	0.998	0.964	0.951	1.018	0.981	0.978	1.056
		Skew	0.265	0.234	0.220	0.296	0.221	0.220	0.220
	Uniform	Mean	-0.026	-0.010	-0.001	-0.021	-0.001	-0.001	
		Median	-0.620	-0.414	-0.313	-0.729	-0.338	-0.323	
		SD	1.052	1.021	1.010	1.048	1.011	1.008	
		Skew	0.259	0.234	0.221	0.285	0.221	0.220	
Normal	Normal	Mean	-0.068	-0.072	-0.074	-0.067	-0.076	-0.076	-0.079
		Median	-0.041	-0.026	-0.017	-0.042	-0.092	-0.014	-0.010
		SD	1.002	0.943	0.924	1.037	0.954	0.949	0.977
		Skew	0.025	0.010	0.006	0.002	0.014	0.002	0.005
	Uniform	Mean	-0.074	-0.077	-0.079	-0.070	-0.079	-0.079	
		Median	-0.041	-0.026	-0.018	-0.042	-0.089	-0.015	
		SD	1.054	1.004	0.987	1.064	0.986	0.980	
		Skew	0.021	0.014	0.012	-0.001	0.015	0.005	
Positive Skew	Normal	Mean	-0.055	-0.010	0.010	-0.060	0.002	0.011	0.014
		Median	-0.361	-0.229	-0.156	-0.488	-0.212	-0.158	-0.156
		SD	0.826	0.745	0.716	0.860	0.740	0.734	0.754
		Skew	0.854	1.056	1.141	1.157	1.156	1.143	1.157
	Uniform	Mean	-0.046	-0.006	0.012	-0.053	0.004	0.012	
		Median	-0.361	-0.233	-0.164	-0.488	-0.215	-0.163	
		SD	0.860	0.790	0.763	0.876	0.764	0.758	
		Skew	0.916	1.094	1.169	0.874	1.168	1.157	
Uniform	Normal	Mean	-0.178	-0.182	-0.185	-0.185	-0.192	-0.193	-0.199
		Median	-0.733	-0.520	-0.429	-0.792	-0.432	-0.430	-0.438
		SD	1.606	1.595	1.591	1.652	1.642	1.641	1.695
		Skew	0.174	0.181	0.184	0.173	0.181	0.182	0.180
	Uniform	Mean	-0.193	-0.197	-0.199	-0.193	-0.200	-0.200	
		Median	-0.734	-0.537	-0.454	-0.792	-0.444	-0.443	
		SD	1.733	1.723	1.717	1.721	1.708	1.706	
		Skew	0.169	0.175	0.178	0.170	0.178	0.178	

Table 2: Fidelity coefficients.

Latent θ Distribution	Prior	Test Length 61 Quadrature Points 10	61 Quadrature Points $2*\sqrt{length}$[a]	61 Quadrature Points 80	Test Length 122 Quadrature Points 10	122 Quadrature Points $2*\sqrt{length}$[a]	122 Quadrature Points 80
Bimodal	Normal	0.9910	0.9990	1.0000	0.9865	0.9998	1.0000
	Uniform	0.9935	0.9993	1.0000	0.9886	0.9998	1.0000
Normal	Normal	0.9835	0.9978	0.9999	0.9765	0.9994	1.0000
	Uniform	0.9877	0.9984	0.9999	0.9793	0.9995	1.0000
Positive Skew	Normal	0.9749	0.9961	0.9999	0.9587	0.9987	0.9999
	Uniform	0.9798	0.9971	0.9999	0.9620	0.9989	0.9999
Uniform	Normal	0.9975	0.9997	1.0000	0.9960	0.9999	1.0000
	Uniform	0.9983	0.9998	1.0000	0.9966	0.9999	1.0000

[a]length= test length

Table 3: Mean EAP, Empirical, and Difference SEEs[a]

Latent θ Distribution	Prior	SEE[a]	Test Length 61 Quadrature Points 10	61 Quadrature Points $2*\sqrt{length}$[b]	61 Quadrature Points 80	Test Length 122 Quadrature Points 10	122 Quadrature Points $2*\sqrt{length}$[b]	122 Quadrature Points 80
Bimodal	Normal	Emp	0.248	0.227	0.232	0.202	0.166	0.169
		EAP	0.160	0.225	0.239	0.093	0.167	0.171
		Diff	0.087	0.003	-0.007	0.108	-0.001	-0.003
	Uniform	Emp	0.263	0.243	0.247	0.209	0.172	0.174
		EAP	0.175	0.234	0.246	0.098	0.170	0.174
		Diff	0.088	0.008	0.001	0.111	0.002	0.000
Normal	Normal	Emp	0.302	0.249	0.234	0.236	0.178	0.170
		EAP	0.155	0.224	0.240	0.081	0.166	0.172
		Diff	0.147	0.025	-0.007	0.155	0.012	-0.002
	Uniform	Emp	0.314	0.263	0.250	0.240	0.184	0.176
		EAP	0.167	0.234	0.248	0.084	0.169	0.175
		Diff	0.147	0.029	0.002	0.157	0.015	0.001
Positive Skew	Normal	Emp	0.343	0.263	0.234	0.268	0.176	0.170
		EAP	0.152	0.222	0.239	0.076	0.162	0.171
		Diff	0.191	0.041	-0.006	0.192	0.013	-0.002
	Uniform	Emp	0.353	0.275	0.249	0.273	0.181	0.175
		EAP	0.162	0.231	0.247	0.079	0.166	0.174
		Diff	0.191	0.044	0.002	0.194	0.015	0.001
Uniform	Normal	Emp	0.275	0.245	0.240	0.234	0.180	0.177
		EAP	0.203	0.244	0.250	0.120	0.178	0.181
		Diff	0.072	0.001	-0.010	0.114	0.002	-0.004
	Uniform	Emp	0.303	0.272	0.266	0.248	0.190	0.187
		EAP	0.224	0.259	0.263	0.130	0.183	0.186
		Diff	0.079	0.013	0.002	0.118	0.007	0.002

[a]Emp = $SEE_{empirical}$, EAP = SEE_{EAP}, Diff = ($SEE_{empirical}$ - SEE_{EAP}) ; [b]length= test length

Table 4: Descriptive statistics on 68% and 95% confidence intervals[a]

Latent θ Distribution	Prior		Test Length 61 Quadrature Points 10	61 Quadrature Points 2*√length[b]	61 Quadrature Points 80	Test Length 122 Quadrature Points 10	122 Quadrature Points 2*√length[b]	122 Quadrature Points 80
68% confidence interval								
Bimodal	Normal	Mean	335.5	663.8	679.4	177.1	675.9	686.9
		SD	214.2	156.0	44.0	132.2	85.8	34.9
	Uniform	Mean	372.6	658.0	684.5	186.0	675.7	686.1
		SD	230.5	143.5	44.2	135.6	75.8	35.4
Normal	Normal	Mean	289.5	582.6	685.1	120.1	609.5	684.8
		SD	224.6	209.0	49.0	118.1	166.0	41.1
	Uniform	Mean	305.2	586.1	684.0	136.4	615.7	677.4
		SD	237.0	205.8	44.3	130.5	166.3	42.8
Positive Skew	Normal	Mean	238.0	540.1	677.6	103.9	608.1	685.9
		SD	203.7	228.6	51.7	134.6	185.0	43.4
	Uniform	Mean	269.9	543.5	672.8	113.8	610.3	680.7
		SD	236.0	221.1	52.6	147.9	186.8	39.9
Uniform	Normal	Mean	434.1	630.0	664.3	242.8	652.3	669.8
		SD	236.8	138.0	57.1	148.1	111.0	40.0
	Uniform	Mean	463.1	649.6	675.8	263.7	654.8	682.8
		SD	234.4	135.3	51.0	153.2	114.4	38.7
95% confidence interval								
Bimodal	Normal	Mean	496.9	904.6	948.8	280.6	934.7	950.2
		SD	248.5	113.1	16.5	187.3	71.6	11.5
	Uniform	Mean	549.6	918.0	950.8	291.0	935.0	950.1
		SD	254.3	105.3	13.0	184.5	71.2	11.0
Normal	Normal	Mean	413.2	844.7	949.6	208.8	868.7	948.7
		SD	271.5	173.5	16.0	181.6	159.3	15.7
	Uniform	Mean	458.0	858.0	948.5	225.1	872.6	948.9
		SD	292.1	166.0	15.7	189.6	153.8	14.7
Positive Skew	Normal	Mean	345.4	788.7	952.3	176.9	881.8	947.8
		SD	243.3	207.6	16.0	178.0	145.4	16.3
	Uniform	Mean	380.0	808.7	948.3	188.3	884.5	946.4
		SD	271.1	201.2	15.3	183.0	144.8	15.5
Uniform	Normal	Mean	634.3	893.7	935.0	383.3	919.8	942.5
		SD	261.6	122.6	27.5	210.0	87.2	15.3
	Uniform	Mean	676.4	913.2	949.3	411.0	925.6	948.1
		SD	268.3	124.3	14.4	209.6	88.4	9.9

[a]SD=standard deviation; [b]length= test length

Table 5: Descriptive statistics on RMSE(θ) and Bias(θ)[a]

Latent θ Distribution	Prior		Test Length 61 Quadrature Points 10	61 Quadrature Points 2*√length[b]	61 Quadrature Points 80	122 Quadrature Points 10	122 Quadrature Points 2*√length[b]	122 Quadrature Points 80
RMSE(θ)								
Bimodal	Normal	Mean	0.282	0.235	0.238	0.265	0.169	0.171
		SD	0.122	0.044	0.009	0.137	0.020	0.006
	Uniform	Mean	0.293	0.246	0.247	0.267	0.173	0.174
		SD	0.114	0.040	0.008	0.135	0.019	0.005
Normal	Normal	Mean	0.352	0.258	0.239	0.328	0.183	0.172
		SD	0.186	0.070	0.016	0.207	0.042	0.011
	Uniform	Mean	0.361	0.270	0.250	0.330	0.187	0.176
		SD	0.177	0.065	0.013	0.204	0.041	0.010
Positive Skew	Normal	Mean	0.400	0.272	0.237	0.379	0.180	0.171
		SD	0.206	0.079	0.013	0.228	0.042	0.010
	Uniform	Mean	0.410	0.284	0.249	0.382	0.185	0.175
		SD	0.195	0.074	0.011	0.225	0.040	0.009
Uniform	Normal	Mean	0.311	0.267	0.261	0.280	0.188	0.185
		SD	0.121	0.051	0.032	0.142	0.029	0.020
	Uniform	Mean	0.321	0.275	0.267	0.284	0.192	0.188
		SD	0.115	0.049	0.028	0.137	0.030	0.021
Bias(θ)								
Bimodal	Normal	Mean	-0.026	-0.011	-0.001	-0.019	-0.001	-0.001
		SD	0.135	0.060	0.055	0.167	0.031	0.028
	Uniform	Mean	-0.026	-0.010	-0.001	-0.021	-0.001	-0.001
		SD	0.127	0.042	0.009	0.161	0.019	0.007
Normal	Normal	Mean	0.011	0.007	0.005	0.012	0.003	0.003
		SD	0.181	0.072	0.054	0.226	0.041	0.029
	Uniform	Mean	0.005	0.002	0.000	0.009	0.000	0.000
		SD	0.177	0.062	0.014	0.225	0.033	0.009
Positive Skew	Normal	Mean	-0.070	-0.024	-0.005	-0.074	-0.012	-0.003
		SD	0.191	0.067	0.040	0.254	0.040	0.023
	Uniform	Mean	-0.060	-0.020	-0.002	-0.067	-0.011	-0.002
		SD	0.193	0.069	0.014	0.255	0.037	0.009
Uniform	Normal	Mean	0.021	0.016	0.014	0.014	0.007	0.006
		SD	0.147	0.108	0.105	0.156	0.057	0.055
	Uniform	Mean	0.006	0.002	0.000	0.006	-0.001	-0.001
		SD	0.108	0.046	0.027	0.142	0.026	0.015

[a]SD=standard deviation; [b]length= test length

Table 6: Repeated Measures Analysis of SEE difference ($SEE_{empirical}$ - $SEE_{asymptotic}$)

Source	v_1	v_2	F
Latent[a]	3	396	13.26**
Length[b]	1	396	5.51*
QuadPts[c]	2	395	716.30**
Prior[d]	1	396	363.67**
Latent X Length	3	396	18.51**
Latent X QuadPts	6	790	8.38**
Latent X Prior	3	396	25.40**
Length X QuadPts	2	395	48.39**
Length X Prior	1	396	188.15**
QuadPts X Prior	2	395	164.55**
Latent X Length X QuadPts	6	790	11.47**
Latent X Length X Prior	3	396	30.05**
Latent X QuadPts X Prior	6	790	14.67**
Length X QuadPts X Prior	2	395	236.59**
Latent X Length X QuadPts X Prior	6	790	16.25**

[a]Latent Distribution; [b]Test Length; [c]Number of Quadrature Points; [d]Prior Distribution; * $p < 0.05$, ** $p < 0.01$

Table 7: Post Hoc Analyses (t-tests) for SEE difference ($SEE_{empirical} - SEE_{EAP}$).

Latent θ Distribution	Hypotheses	Prior Distribution: Normal, Test Length 61	Normal, Test Length 122	Uniform, Test Length 61	Uniform, Test Length 122
Bimodal	μ_{10} vs $\mu_{2*\sqrt{length}}$	12.58**	16.16**	11.82**	16.25**
	μ_{10} vs μ_{80}	13.94**	16.49**	12.93**	16.53**
	$\mu_{2*\sqrt{length}}$ vs μ_{80}	1.36	0.32	1.11	0.28
Normal	μ_{10} vs $\mu_{2*\sqrt{length}}$	11.35**	13.20**	10.96**	13.23**
	μ_{10} vs μ_{80}	14.27**	14.58**	13.56**	14.54**
	$\mu_{2*\sqrt{length}}$ vs μ_{80}	2.92**	1.37	2.59**	1.30
Positive Skew	μ_{10} vs $\mu_{2*\sqrt{length}}$	12.28**	14.70**	12.17**	14.85**
	μ_{10} vs μ_{80}	16.15**	15.92**	15.68**	16.01**
	$\mu_{2*\sqrt{length}}$ vs μ_{80}	3.86**	1.22	3.51**	1.16
Uniform	μ_{10} vs $\mu_{2*\sqrt{length}}$	9.53**	14.98**	8.99**	15.08**
	μ_{10} vs μ_{80}	11.05**	15.72**	10.45**	15.80**
	$\mu_{2*\sqrt{length}}$ vs μ_{80}	1.52	0.74	1.46	0.72

Prior Distribution	Quadrature Points[a]	Hypotheses	Latent Distribution: Bimodal	Normal	Positive Skew	Uniform
Normal	10	μ_{61} vs μ_{122}	-6.01**	-1.83	-0.28	-12.39**
	$2*\sqrt{length}$		0.91	3.03**	6.40**	-0.26
	80		-1.10	-1.02	-0.91	-1.98*
Uniform	10	μ_{61} vs μ_{122}	-6.89**	-2.43**	-0.78	-12.37**
	$2*\sqrt{length}$		1.98*	3.75**	6.84**	2.07*
	80		0.31	0.24	0.16	0.31

[a]length= test length; * $p < 0.05$, ** $p < 0.01$

Table 8: Repeated Measures Analysis of diff68% and diff95%.

Source	v_1	v_2	$F_{diff68\%}$	$F_{diff95\%}$
Latent[a]	3	396	21.41**	28.70**
Length[b]	1	396	149.38**	288.73**
QuadPts[c]	2	395	1134.64**	1339.62**
Prior[d]	1	396	34.76**	209.88**
Latent X Length	3	396	4.28**	7.57**
Latent X QuadPts	6	790	12.38**	15.84**
Latent X Prior	3	396	2.71*	3.65*
Length X QuadPts	2	395	263.00**	443.28**
Length X Prior	1	396	6.17*	71.18**
QuadPts X Prior	2	395	26.14**	82.90**
Latent X Length X QuadPts	6	790	2.17*	5.10**
Latent X Length X Prior	3	396	0.51	1.04
Latent X QuadPts X Prior	6	790	1.72	2.14*
Length X QuadPts X Prior	2	395	1.95	27.91**
Latent X Length X QuadPts X Prior	6	790	2.08	2.15*

[a]Latent Distribution; [b]Test Length; [c]Number of Quadrature Points; [d]Prior Distribution; * $p < 0.05$, ** $p < 0.01$

Table 9: Post Hoc Analyses (t-tests) for diff68% (CI68% - 680).

Latent Ability Distribution X Prior Distribution

Hypotheses[a]	Prior Distribution: Normal	Uniform
μ_{bi} vs μ_{nml}	3.65**	3.81**
μ_{bi} vs μ_{ps}	-5.39**	-5.49**
μ_{bi} vs μ_{unif}	-1.10	-1.87
μ_{nml} vs μ_{ps}	-1.74	-1.68
μ_{nml} vs μ_{unif}	-4.75**	-5.69**
μ_{ps} vs μ_{unif}	-6.49**	-7.37**

Test Length X Prior Distribution

Hypotheses[a]	Prior Distribution: Normal	Uniform
μ_{61} vs μ_{122}	10.67**	12.34**

	Test Length: 61	122
μ_{nml} vs μ_{unif}	-6.06**	-2.77**

Quadrature Points X Prior Distribution

Hypotheses[b]	Prior Distribution: Normal	Uniform
μ_{10} vs $\mu_{2*\sqrt{length}}$	-42.24**	-40.31**
μ_{10} vs μ_{80}	-48.83**	-46.61**
$\mu_{2*\sqrt{length}}$ vs μ_{80}	-6.59**	-6.30**

Hypothesis[a]	Quadrature Points: 10	$2*\sqrt{\text{test length}}$[a]	80
μ_{nml} vs μ_{unif}	-9.86**	-1.82	-0.62

Quadrature Points X Test Length within Latent Ability Distribution

Latent θ Distribution	Hypotheses[b]	Test Length: 61	122	Hypotheses	Quadrature Points: 10	$2*\sqrt{\text{test length}}$[b]	80
Bimodal	μ_{10} vs $\mu_{2*\sqrt{length}}$	-28.16**	-45.34**	μ_{61} vs μ_{122}	34.00**	-2.92**	0.89
	μ_{10} vs μ_{80}	-30.09**	-46.32**				
	$\mu_{2*\sqrt{length}}$ vs μ_{80}	-1.93	-0.98				
Normal	μ_{10} vs $\mu_{2*\sqrt{length}}$	-20.92**	-35.31**	μ_{61} vs μ_{122}	37.83**	-6.32**	0.77
	μ_{10} vs μ_{80}	-28.22**	-40.30**				
	$\mu_{2*\sqrt{length}}$ vs μ_{80}	-7.30**	-4.99**				
Positive Skew	μ_{10} vs $\mu_{2*\sqrt{length}}$	-21.10**	-36.69**	μ_{61} vs μ_{122}	23.32**	-10.83**	-1.31
	μ_{10} vs μ_{80}	-30.88**	-42.12**				
	$\mu_{2*\sqrt{length}}$ vs μ_{80}	-9.78**	-5.43**				
Uniform	μ_{10} vs $\mu_{2*\sqrt{length}}$	-16.84**	-35.24**	μ_{61} vs μ_{122}	39.50**	-2.77**	-1.26
	μ_{10} vs μ_{80}	-19.50**	-37.24**				
	$\mu_{2*\sqrt{lengtn}}$ vs μ_{80}	-2.66**	-2.00*				

[a]bi=bimodal, nml=normal, ps=positive skew, unif=uniform; [b]length= test length;
* $p < 0.05$, ** $p < 0.01$

Table 10: Post Hoc Analyses (t-tests) for diff95% (CI95% - 950).

Latent θ Distribution	Hypotheses[a]	Prior Distribution: Normal, Test Length 61	Normal, Test Length 122	Uniform, Test Length 61	Uniform, Test Length 122
Bimodal	μ_{10} vs $\mu_2*\sqrt{length}$	-21.11**	-33.87**	-19.43**	-33.95**
	μ_{10} vs μ_{80}	-23.40**	-34.67**	-21.16**	-34.75**
	$\mu_2*\sqrt{length}$ vs μ_{80}	-2.29*	-0.80	-1.73	-0.80
Normal	μ_{10} vs $\mu_2*\sqrt{length}$	-19.14**	-29.27**	-17.78**	-28.79**
	μ_{10} vs μ_{80}	-23.78**	-32.81**	-21.80**	-32.18**
	$\mu_2*\sqrt{length}$ vs μ_{80}	-4.65**	-3.54**	-4.02**	-3.39**
Positive Skew	μ_{10} vs $\mu_2*\sqrt{length}$	20.17**	-32.07**	-19.08**	-30.99**
	μ_{10} vs μ_{80}	-27.61**	-35.07**	-25.30**	-33.74**
	$\mu_2*\sqrt{length}$ vs μ_{80}	7.44**	-3.00**	-6.22**	-2.76**
Uniform	μ_{10} vs $\mu_2*\sqrt{length}$	-12.60**	-26.05**	-11.73**	-25.50**
	μ_{10} vs μ_{80}	-14.60**	-27.15**	-13.52**	-26.62**
	$\mu_2*\sqrt{length}$ vs μ_{80}	-2.00*	-1.10	-1.79	-1.12

Prior Distribution	Quadrature Points[a]	Hypotheses	Latent Distribution: Bimodal	Normal	Positive Skew	Uniform
Normal	10	μ_{61} vs μ_{122}	20.92**	17.26	11.92**	21.21**
	$2*\sqrt{length}$		-2.91**	-2.03*	-6.58**	-2.20*
	80		-0.13	0.08	0.32	-0.64
Uniform	10	μ_{61} vs μ_{122}	24.54**	18.15**	13.10**	23.65**
	$2*\sqrt{length}$		-1.61	-1.14	-5.18**	-1.11
	80		0.07	-0.03	0.13	0.10

[a]length= test length; * $p < 0.05$, ** $p < 0.01$

Table 11: Repeated Measures Analysis of RMSE(θ) and Bias(θ).

Source	v_1	v_2	F_{RMSE}	F_{Bias}
Latent[a]	3	396	7.70**	5.39**
Length[b]	1	396	4281.22**	0.12
QuadPts[c]	2	395	176.33**	3.18*
Prior[d]	1	396	526.39**	0.78
Latent X Length	3	396	6.91**	3.44*
Latent X QuadPts	6	790	6.44**	2.35*
Latent X Prior	3	396	1.28	1.12
Length X QuadPts	2	395	302.98**	3.37*
Length X Prior	1	396	687.96**	0.55
QuadPts X Prior	2	395	17.35**	2.71
Latent X Length X QuadPts	6	790	4.14**	2.00
Latent X Length X Prior	3	396	8.53**	1.15
Latent X QuadPts X Prior	6	790	2.21*	3.77**
Length X QuadPts X Prior	2	395	12.91**	2.65
Latent X Length X QuadPts X Prior	6	790	1.69	2.37*

[a]Latent Distribution; [b]Test Length; [c]Number of Quadrature Points; [d]Prior Distribution;
* $p < 0.05$, ** $p < 0.01$

Table 12: Post Hoc Analyses (t-tests) for RMSE(θ).

Quadrature Points X Prior Distribution within Latent Ability Distribution

	Latent θ Distribution[b]			
Hypotheses[a,b]	bi	nml	ps	unif
μ_{nml} vs μ_{unif}	-17.13**	-17.06**	-24.50**	-11.16**
μ_{10} vs $\mu_{2*\sqrt{length}}$	6.36**	10.60**	14.51**	6.12**
μ_{10} vs m_{80}	6.19**	11.95**	16.51**	6.63**
$\mu_{2*\sqrt{length}}$ vs μ_{80}	-0.17	1.35	2.00*	0.51

Quadrature Points X Test Length within Latent Ability Distribution

Latent θ		Test Length			Quadrature Points		
Distribution	Hypotheses[a]	61	122	Hypotheses	10	$2*\sqrt{\text{test length}}$[a]	80
Bimodal	μ_{10} vs $\mu_{2*\sqrt{length}}$	6.99**	14.11**	μ_{61} vs μ_{122}	18.09**	58.52**	59.37**
	μ_{10} vs μ_{80}	6.64**	13.91**				
	$\mu_{2*\sqrt{length}}$ vs μ_{80}	-0.35	-0.20				
Normal	μ_{10} vs $\mu_{2*\sqrt{length}}$	9.15**	14.36**	μ_{61} vs μ_{122}	24.75**	72.45**	64.17**
	μ_{10} vs μ_{80}	11.11**	15.41**				
	$\mu_{2*\sqrt{length}}$ vs μ_{80}	1.95	1.05				
Positive	μ_{10} vs $\mu_{2*\sqrt{length}}$	11.31**	17.70**	μ_{61} vs μ_{122}	17.72**	69.26**	50.29**
Skew	μ_{10} vs μ_{80}	14.49**	18.53**				
	$\mu_{2*\sqrt{length}}$ vs μ_{80}	3.18**	0.83				
Uniform	μ_{10} vs $\mu_{2*\sqrt{length}}$	6.53**	13.37**	μ_{61} vs μ_{122}	23.10**	54.80**	52.40**
	μ_{10} vs μ_{80}	7.62**	13.94**				
	$\mu_{2*\sqrt{length}}$ vs μ_{80}	1.09	0.57				

Prior Distribution X Test Length within Latent Ability Distribution

	Latent θ Distribution[b]			
Hypotheses[b]	bi	nml	ps	unif
μ_{nml} vs μ_{unif}	-12.08**	-12.03**	-17.28**	-7.86**
μ_{61} vs μ_{122}	31.97**	37.94**	32.27**	30.64**

Table 12: Post Hoc Analyses (t-tests) for RMSE(θ) (continued).

Quadrature Points X Test Length within Prior Distribution

Prior		Quadrature Points				Test Length	
Distribution	Hypotheses	10	$2*\sqrt{\text{test length}}$[a]	80	Hypotheses[a]	61	122
Normal	μ_{61} vs μ_{122}	23.97**	79.45**	70.41**	μ_{10} vs $\mu_{2*\sqrt{\text{length}}}$	8.28**	14.03**
					μ_{10} vs μ_{80}	9.79**	14.61**
					$\mu_{2*\sqrt{\text{length}}}$ vs μ_{80}	1.52	0.58
Uniform	μ_{61} vs μ_{122}	30.79**	86.80**	76.50**	μ_{10} vs $\mu_{2*\sqrt{\text{length}}}$	8.13**	13.94**
					μ_{10} vs μ_{80}	9.80**	14.54**
					$\mu_{2*\sqrt{\text{length}}}$ vs μ_{80}	1.67	0.61

[a]length= test length; [b]bi=bimodal, nml=normal, ps=positive skew, unif=uniform; * $p < 0.05$, ** $p < 0.01$

Table 13: Post Hoc Analyses (t-tests) for Bias(θ).

Latent θ Distribution	Hypotheses[a]	Prior Distribution: Normal, Test Length 61	Normal, Test Length 122	Uniform, Test Length 61	Uniform, Test Length 122
Bimodal	μ_{10} vs $\mu_{2*\sqrt{length}}$	-1.01	0.07	-0.68	0.01
	μ_{10} vs μ_{80}	-2.08	0.07	-2.44*	0.06
	$\mu_{2*\sqrt{length}}$ vs μ_{80}	-1.06	0.01	-1.76	0.05
Normal	μ_{10} vs $\mu_{2*\sqrt{length}}$	-0.10	-0.24	-0.41	0.00
	μ_{10} vs μ_{80}	0.42	-0.01	0.35	-0.00
	$\mu_{2*\sqrt{length}}$ vs μ_{80}	0.52	0.24	0.76	-0.01
Positive Skew	μ_{10} vs $\mu_{2*\sqrt{length}}$	0.42	-0.71	0.73	-0.83
	μ_{10} vs μ_{80}	-4.31**	-0.85	-3.99**	-0.86
	$\mu_{2*\sqrt{length}}$ vs μ_{80}	-4.73**	-0.14	-4.72**	-0.02
Uniform	μ_{10} vs $\mu_{2*\sqrt{length}}$	1.00	-1.00	0.04	-0.14
	μ_{10} vs μ_{80}	0.67	0.10	0.70	0.11
	$\mu_{2*\sqrt{length}}$ vs μ_{80}	-0.33	1.10	0.65	0.25

Prior Distribution	Quadrature Points[a]	Hypotheses	Latent Distribution: Bimodal	Normal	Positive Skew	Uniform
Normal	10	μ_{61} vs μ_{122}	-2.05*	0.49	-3.10**	1.26
	$2*\sqrt{length}$		-1.40	0.40	-3.74**	-0.03
	80		-0.76	0.24	-1.12	0.90
Uniform	10	μ_{61} vs μ_{122}	-2.19*	0.30	-2.78**	0.69
	$2*\sqrt{length}$		-1.79	0.53	-3.66**	0.57
	80		-0.74	0.10	-1.02	0.33

[a]length= test length; * $p < 0.05$, ** $p < 0.01$

Table 14: Comparison of RMSEs based on EAP SEE and Empirical SEE with Observed RMSE.

Latent θ Distribution	Test Length	Quadrature Points[a]	Prior Distribution: Normal Emp[b]	Normal EAP	Normal Observed	Uniform Emp[b]	Uniform EAP	Uniform Observed
Bimodal	61	10	0.249	0.162	0.282	0.264	0.177	0.293
		$2*\sqrt{\text{test length}}$	0.227	0.225	0.235	0.243	0.234	0.246
		80	0.232	0.239	0.238	0.247	0.246	0.247
	122	10	0.203	0.095	0.265	0.210	0.092	0.267
		$2*\sqrt{\text{test length}}$	0.166	0.167	0.169	0.172	0.170	0.173
		80	0.169	0.171	0.171	0.174	0.174	0.174
Normal	61	10	0.302	0.155	0.352	0.314	0.167	0.361
		$2*\sqrt{\text{test length}}$	0.249	0.224	0.258	0.263	0.234	0.270
		80	0.234	0.240	0.239	0.250	0.248	0.250
	122	10	0.236	0.082	0.328	0.240	0.084	0.330
		$2*\sqrt{\text{test length}}$	0.178	0.166	0.183	0.184	0.169	0.187
		80	0.170	0.172	0.172	0.176	0.175	0.176
Positive Skew	61	10	0.350	0.167	0.400	0.358	0.173	0.410
		$2*\sqrt{\text{test length}}$	0.264	0.223	0.272	0.276	0.232	0.284
		80	0.234	0.239	0.237	0.249	0.247	0.249
	122	10	0.278	0.106	0.379	0.281	0.104	0.382
		$2*\sqrt{\text{test length}}$	0.176	0.162	0.180	0.181	0.166	0.185
		80	0.170	0.171	0.171	0.175	0.174	0.175
Uniform	61	10	0.276	0.204	0.311	0.303	0.224	0.321
		$2*\sqrt{\text{test length}}$	0.246	0.245	0.267	0.272	0.259	0.275
		80	0.240	0.250	0.261	0.266	0.263	0.267
	122	10	0.234	0.121	0.280	0.248	0.130	0.284
		$2*\sqrt{\text{test length}}$	0.180	0.178	0.188	0.190	0.183	0.192
		80	0.177	0.181	0.185	0.187	0.186	0.188

[a]length= test length; [b]Emp=Empirical

Figure Captions

Figure 1a. Mean SEEs for Test Length X Prior Distribution X Number of Quadrature Points Interaction for Bimodal θ distribution 61-item test length.

Figure 1b. Mean SEEs for Test Length X Prior Distribution X Number of Quadrature Points Interaction for Bimodal θ distribution for 122-item test length.

Figure 1c. Mean SEEs for Test Length X Prior Distribution X Number of Quadrature Points Interaction for Normal θ distribution 61-item test length.

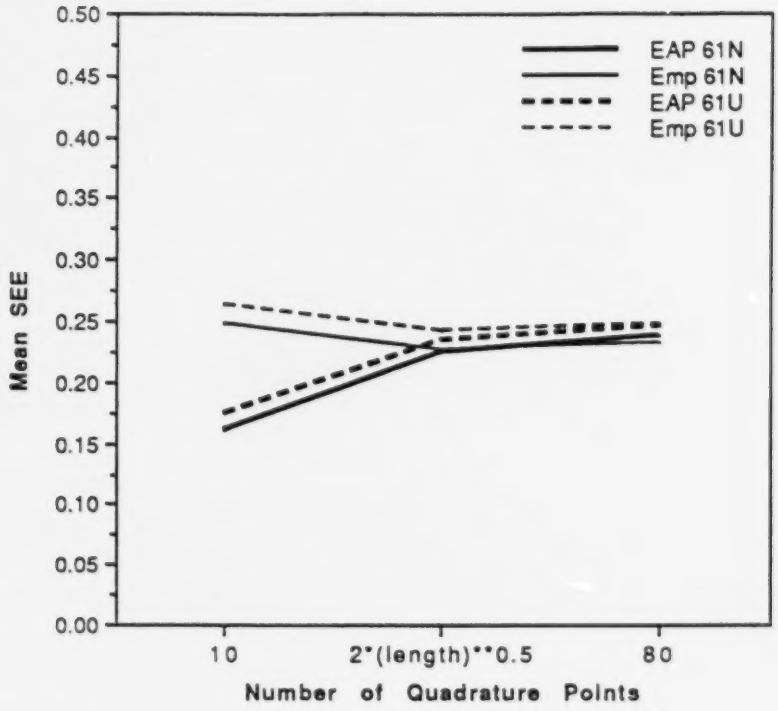
Figure 1d. Mean SEEs for Test Length X Prior Distribution X Number of Quadrature Points Interaction for Normal θ distribution for 122-item test length.

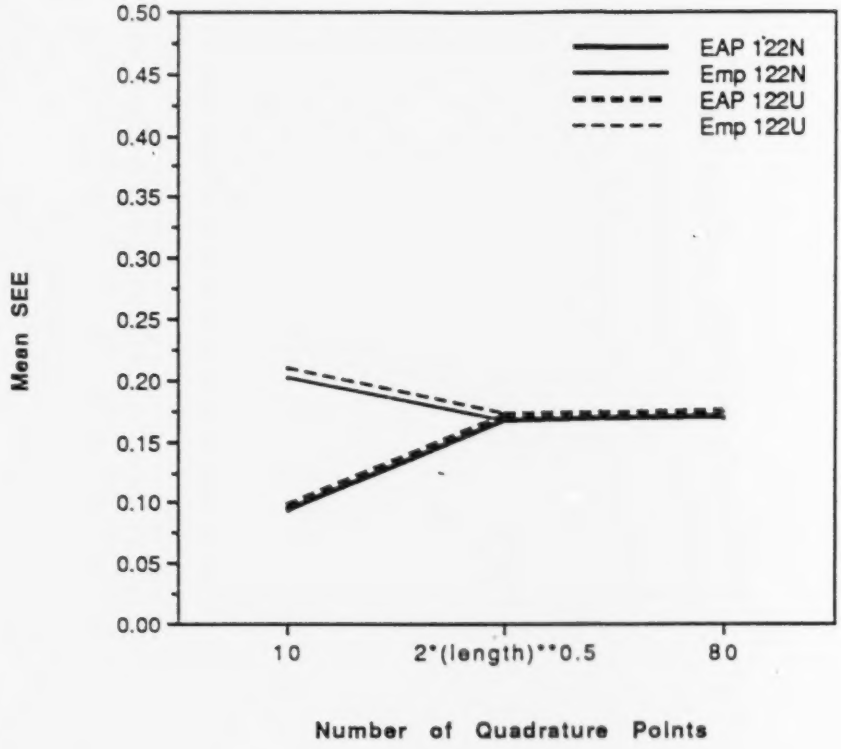
Figure 1e. Mean SEEs for Test Length X Prior Distribution X Number of Quadrature Points Interaction for Positive Skew θ distribution for 61-item test length.

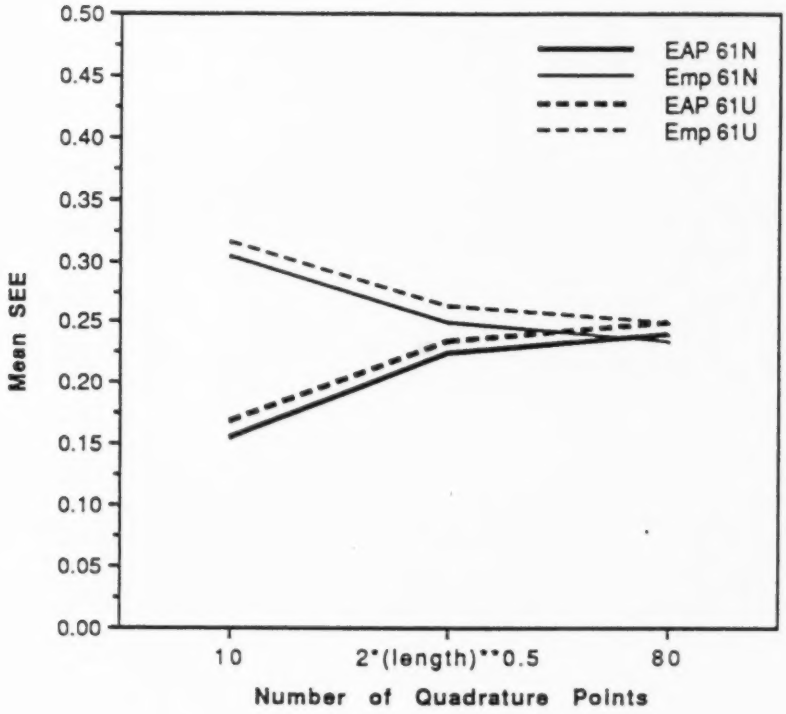
Figure 1f. Mean SEEs for Test Length X Prior Distribution X Number of Quadrature Points Interaction for Positive Skew θ distribution for 122-item test length.

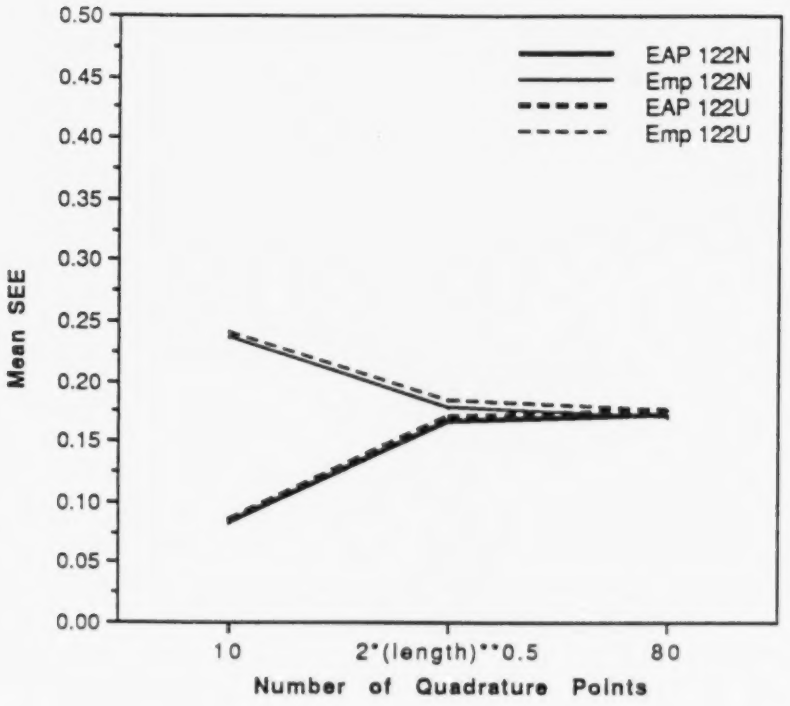
Figure 1g. Mean SEEs for Test Length X Prior Distribution X Number of Quadrature Points Interaction for Uniform θ distribution 61-item test length.

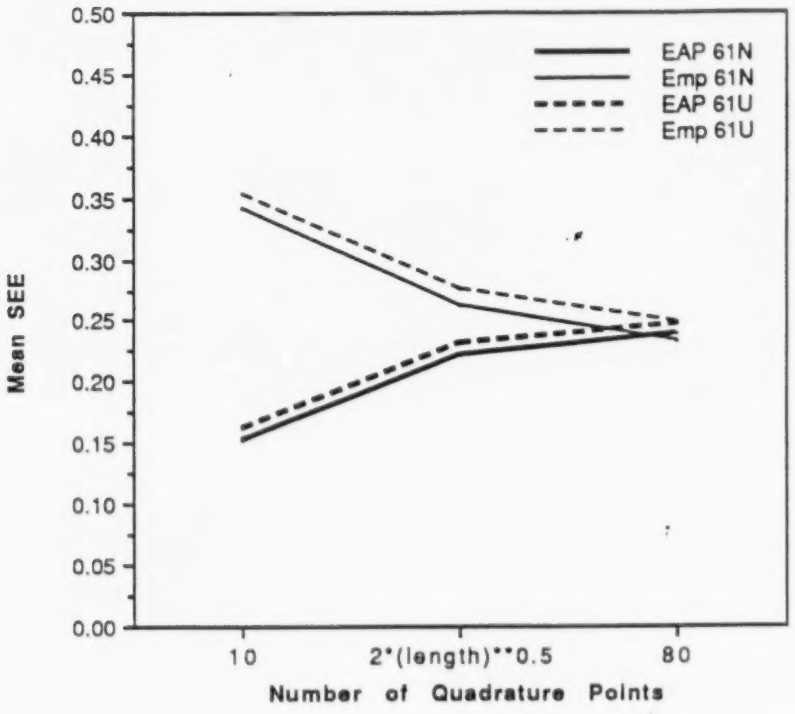
Figure 1h. Mean SEEs for Test Length X Prior Distribution X Number of Quadrature Points Interaction for Uniform θ distribution for 122-item test length.

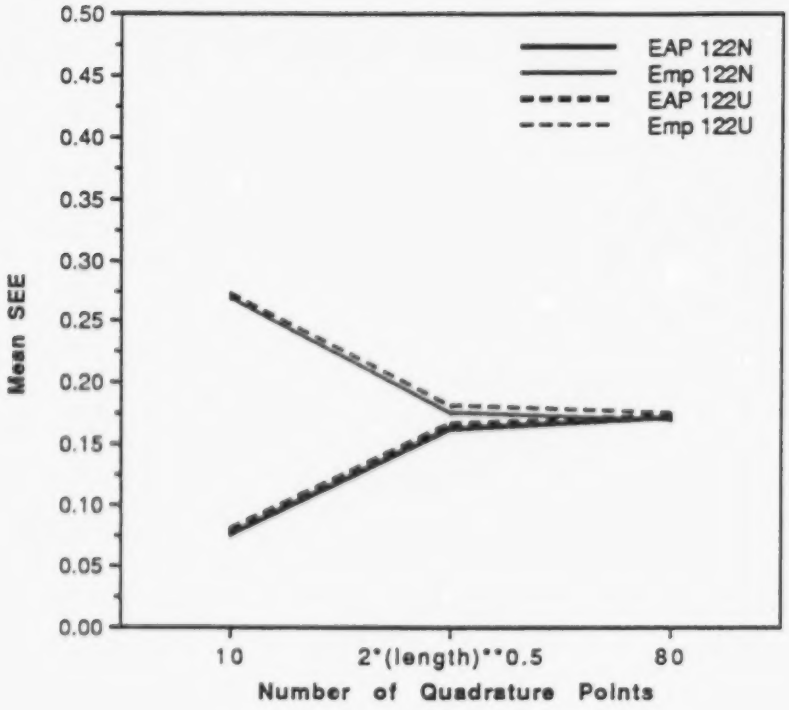

0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Mean SEE
EAP 61N
Emp 61N
EAP 61U
Emp 61U
10
2*(length)**0.5
80
Number of Quadrature Points

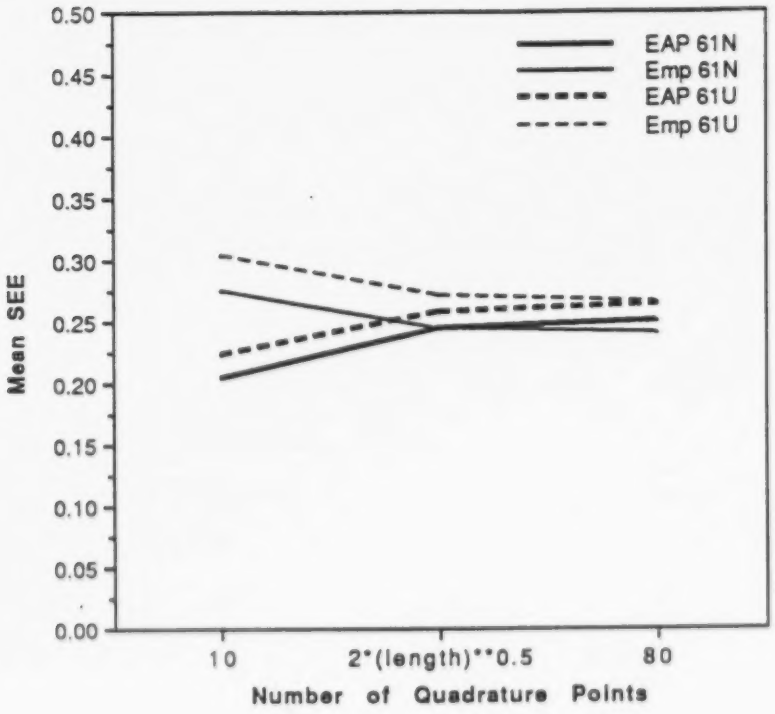

0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Mean SEE
EAP 122N
Emp 122N
EAP 122U
Emp 122U
10
2*(length)**0.5
80
Number of Quadrature Points

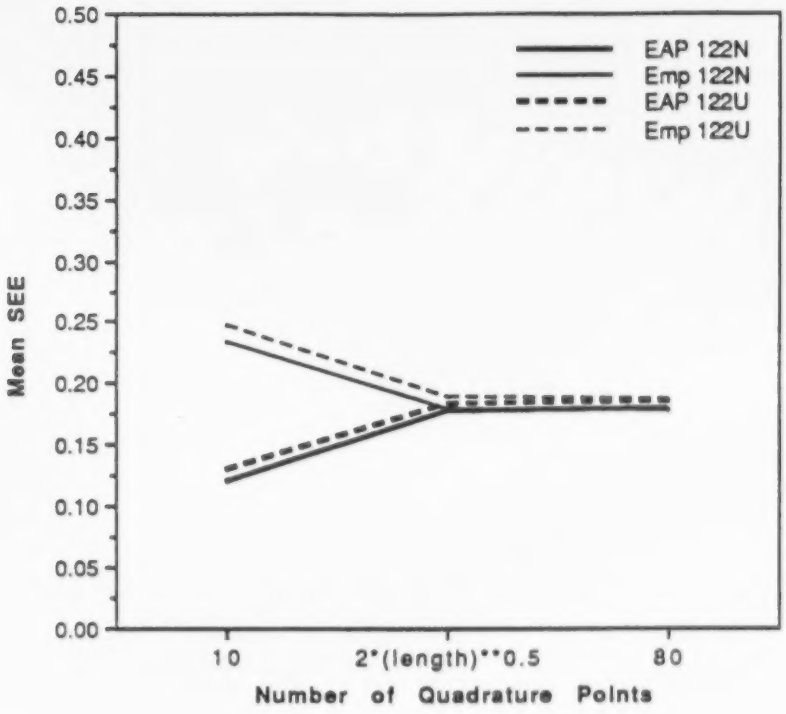

0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Mean SEE
EAP 61N
Emp 61N
EAP 61U
Emp 61U
10
2*(length)**0.5
80
Number of Quadrature Points


0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Mean SEE
EAP 122N
Emp 122N
EAP 122U
Emp 122U
10
2*(length)**0.5
80
Number of Quadrature Points


0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Mean SEE
EAP 61N
Emp 61N
EAP 61U
Emp 61U
10
2*(length)**0.5
80
Number of Quadrature Points


0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Mean SEE
EAP 122N
Emp 122N
EAP 122U
Emp 122U
10
2*(length)**0.5
80
Number of Quadrature Points


0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Mean SEE
EAP 61N
Emp 61N
EAP 61U
Emp 61U
10
2*(length)**0.5
80
Number of Quadrature Points


0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Mean SEE
EAP 122N
Emp 122N
EAP 122U
Emp 122U
10
2*(length)**0.5
80
Number of Quadrature Points